機械学習による代名詞「自分」の人称判別システム

長崎　英紀*　　古宮　嘉那子*　　但馬　康宏*　　小谷　善行*
*東京農工大学大学院　情報工学専攻

1　はじめに

照応解析の対象の１つである代名詞「自分」は，文の意味によって照応する人物・人称が自在に変化するので，解析が困難である．言語学的な立場から代名詞「自分」に焦点を当てた研究が行われており[1]，自然言語処理の解析対象としても，代名詞「自分」は重要である．

また，日本語を他の言語に訳す際，「自分」と同様の働きをする代名詞がその言語に存在しない場合がある．英語を例にとると，日英翻訳した際に「自分」を人称代名詞（I, you, he, himself, one 等）に書き換えなければならない[2]．照応解析・機械翻訳の側面から，代名詞「自分」の人称判別および照応解析は必要であるといえる．

代名詞「自分」の人称が判別できれば，「私」などの人称代名詞と同じ解析方法を，照応解析および機械翻訳に適用できる．本研究は代名詞「自分」の人称を自動で判別するシステムを実装した．

2　「自分」を分類する人称の定義

代名詞「自分」がどの人称を表すのか判断するために，分類する人称の種類を定義する必要がある．最終的に文章内の人物と「自分」の照応関係を求めるために人称を利用するという立場から，人称の種類を定義した．分類した人称の種類は，“人一般”，“話し手”，“聞き手”，“第三者”の 4 種類である．

2.1　特定の人物を指すか

まず，特定の人物を指すか否かで，先行詞の有無も決まる．特定の人物を指さない場合を“人一般”とし，他と区別した．以下に“人一般”の定義をする．

人称定義 1

その文が発せられた時点で，「自分」を指す人物が特定できない　⇒　“人一般”

例 1 では，「アップする人」が「自分」の指す人物と推測されるが，この文が発せられた時点でその人物は不定であるので，“人一般”となる．

（例 1）このサイトでは、自分の顔写真をアップすると、それを使って企業のコマーシャルを作ってくれる。　→　“人一般”

2.2　特定の人物の人称

特定の人物を指す場合，その文に対して“話し手”，“聞き手”，“第三者”の３種類に分類し，以下のように定義した．

人称定義 2-1

・　その１文を発した者　⇒“話し手” ・　その１文を受けた者　⇒“聞き手” ・　“話し手”，“聞き手”以外の特定の人物または集団　⇒“第三者”

（例 2）「僕は自分を安売りしてない。」
→　“話し手”

（例 3）「自分の子供は普通だと思いますか」
→　“聞き手”

（例 4）チャールズが、放送で自分のノーハウを見せてくれた。
→　“第三者”

人称定義 2-2

・　“話し手”を含む　⇒　“話し手” （“話し手”＋“聞き手”，“話し手”＋“第三者”） ・　“話し手”を含まず，“聞き手”を含む ⇒“聞き手” （“聞き手”＋“第三者”）

（例 5）自分の子供でも他人の子供でも、注意をするときは勇気を持っていこうよ！
→　“話し手”

2.3　人称を決める上での判断基準

以上の定義に従って，人手によって正解の人称を決めた．その際に，分類が曖昧な例に対しても一貫性を保つため，判断基準を設けた．

判断基準 1

・1 文は鉤括弧「」，または，句点。で区切られているものとする。

（例 6）木村は「自分の役割がみえてきた」と収穫を口にした。
→　“話し手”

判断基準 2

「自分」を他の代名詞に置き換えたとき，文の意味が変わらない （「私」，「私自身」，「あなた」，「あなた自身」，「彼」，「彼自身」） ⇒　その人称が正解

（例 7）「自分なりに頑張りたい」＝「私なりに頑張りたい」　→　“話し手”

判断基準 3

主語を補ったとき,「自分」が主語と同じ人物を指す ⇒　主語と同じ人称が正解

（例 8）「（私は）自分でも楽しみ」　→“話し手”

判断基準 4

英訳した際に該当する人称代名詞の人称を正解とする one, oneself, you(総称的な意味の), yourself ⇒　人一般 I, myself　⇒　話し手 you, yourself　⇒　聞き手 he, himself　⇒　第三者

3　決定木学習を用いたルールの抽出

手法として,「自分」を含む 1 文内の情報を属性として機械学習（決定木学習）に利用し，解析器を作成した．これにより，言語的要因から人称判別の法則を自動的に検出することができる．決定木学習は生成された決定木を人間が見て，属性の関係を把握し，改良できるという利点があるので採用した.

4　学習に利用した言語的要因

4.1　主語の人称

「自分」が文の主語ではないとき，その人称が主語の人称と一致する傾向がある（例 9）．この規則を見つけ出すのに必要な情報は以下の２つである.

(1)「自分」の表層格（助詞）

(2) 主語の人称

主語は，付随する助詞が「が」「は」「とも」「も」「には」の名詞とした.

（例 9）患者が医師からもらう薬を自分で決める。

4.2　述語の活用形

文の述語の活用形によって,「自分」の人称が限定される場合がある．以下の例 10 のように主語が省略されており述語が命令形の場合,「自分」の人称は“聞き手”となる.

（例 10）「自分のブログくらい自分で考えて書けよ！」

4.2.1　単文内の述語

「自分」を含む単文内の述語（動詞，形容詞，形容動詞，助動詞）の品詞・活用形・活用型を学習させた（表１）.

4.2.2　複文および重文への対応

複文や重文の場合，文全体としての形式を決めるのは,文末にある述語の活用形である（例 11）．単文内の述語同様に，文全体の述語の品詞・活用形・活用型を学習させた（表１）.

（例 11）ポールさんが自分の名前を商標登録しようと動き出した。

表 1　例 11 の述語情報

	属性名	属性値
単文内の述語	先頭形態素「しよ」の品詞	動詞-自立
	活用型	サ変・スル
	活用形	未然ウ接続
	末尾形態素「う」の品詞	助動詞
	活用型	不変化型
	活用形	基本形
文全体の述語	第 1 形態素「動き出し」の品詞	動詞-自立
	活用型	五段・サ行
	活用形	連用形
	第 2 形態素「た」の品詞	助動詞
	活用型	特殊・タ
	活用形	基本形
	第 3～第 4 形態素	NULL
	末尾形態素:第 2 形態素と同じ	-

4.3　会話文と地の文

「自分」が主語だった場合，文の種類によって人称に偏りが見られた（例 12）．文が会話文である場合,「自分」は“話し手”になる傾向があり，地の文では“話し手”以外になる傾向があったので，文の種類も要因の一つとして利用した.

（例 12）「自分が高校の時は、印象に残っていません。」　→　会話文

4.4　その他

以上,「自分」の人称判別に有効であると推測される要因を述べたが，他にも「自分」の係り先が述語でなかった場合を考慮し,「自分」の係り先の品詞，活用形，活用型を学習させた.

（例 13）詞、曲からジャケットデザインまで自分の意見を反映させた。

5　実装方法および実験結果

5.1　実装方法

文から学習に用いる属性の情報を取り出す際，形態素解析に ChaSen[3]，係り受け解析に CaboCha[4]を使用した．機械学習のアルゴリズムには，決定木学習アルゴリズム C4.5 を実装した[5]．

学習に用いた属性は 4 章で述べた言語的要因，計 34 種類である．表 2 に詳細な属性のラベルを示す．なお，１つの単語が複数の形態素から構成されている場合，全ての形態素，または，重要だと考えられる形態素を抜粋して入力した．

表 2　学習に用いた属性

属性の種類	
「自分」につく助詞の情報（助詞が連続するのは 2 つまでと仮定）	第 1 形態素の表記
	品詞
	第 2 形態素の表記
	品詞
主語の情報	表記
	品詞
	人称
「自分」の係り先文節の情報	先頭形態素の品詞
	活用型
	活用形
	末尾形態素の品詞
	活用型
	活用形
単文の述語の情報	先頭形態素の品詞
	活用型
	活用形
	末尾形態素の品詞
	活用型
	活用形
文の種類	文の種類
文全体の述語の情報	第 1～第 2 形態素の品詞
	活用型
	活用形
	末尾形態素の品詞
	活用型
	活用形
計	34 個

データはインターネット上のニュース記事から「自分」を含む文をランダムに 395 件収集した．正解は，文脈を考慮した上で，２章で述べた定義に従い人手によってつけた．表 3 に正解データにおける人称の割合を示す．

表 3　「自分」の正解人称の割合

人称	該当データ（件数）
人一般	121
話し手	154
聞き手	14
第三者	104
不明	2
合計	395

5.2　実験結果

生成された決定木を図 1 に示す．結果は適合率が Open データ（５分割交差検定）で平均 56.7%，Closed データで 99.7%だった．ベースラインは，全てのデータに対し，最も割合の多い“話し手”と答えた場合の 39.8%である．

人称の分類別の正解率を表 4 に示す．この正解率は，人手でつけた正解データに対し，システムが正答した割合である．

表 4　人称別正解率

	正答数	総数	正解率
人一般	51	121	42.1%
話し手	104	154	67.5%
聞き手	3	14	21.4%
第三者	66	104	63.5%
不明	0	2	0.0%
計	224	395	56.7%

6　考察

生成された決定木（図１）から，主語の人称と「自分」の人称が一致するという法則が強く働くことが分かった．文のタイプが会話文であることも重要な要因である．他の属性は該当データが少なかったため法則性が見られなかったが，これは学習データの総数自体が足りないことが原因として考えられる．

データ数以外に精度が低かった原因として，多数の文で主語が省略（ゼロ代名詞化）されていたことが挙げられる．前後の文の主題・焦点などを用いて対策を行う必要がある．

また，人間にもシステムと同じ 1 文の情報だけ与えた場合，どの程度の精度が得られるのか調べるため，アンケートを実施した．対象は 20 代から 50 代の男女 10 人．25 問のアンケートを 2 種類用意し，解答させた．結果は平均適合率 77%であった．残りの 23%の文を人間は文脈を手がかりに

回答していると考えられる．

この精度（77%）は同じ条件下におけるシステムの目標として設定できる．同じアンケートの問題をシステムに解析させたところ，適合率 66%だった．

```
主語人称=聞き手
|<-yes---
| “聞き手” 1件
|---no->
  主語人称=第三者
  |<-yes---
  | “第三者” 15件
  |---no->
    述語(単文)-先頭の活用型=五段・サ行
    |<-yes---
    | 述語(単文)-先頭の活用形=基本形
    | |<-yes---
    | | “人一般” 1件
    | |---no->
    |   “聞き手” 1件
    |---no->
      文のタイプ=会話文
      |<-yes---
      | 自分係り先-先頭の品詞=接頭詞-名詞接続
      | |<-yes---
      | | “人一般” 1件
      | |---no->
      |   述語(単文)-先頭の品詞=形容詞-自立
      |          ⋮
      |---no->
        1文の述語-末尾の品詞=助動詞
        |<-yes---
        | 自分係り先-先頭の品詞=名詞-形容動詞語幹
        |---no->
          “人一般” 10件
                 ⋮
```

図 1　生成された決定木の上部

一方，解答者の答えが全員一致で正解だった割合は 48%(24/50)だった．本システムの精度はこの数値を上回っているため，解答者全員の答えが一致するような基本的な問題に対して，対応できたと言える．

参考までに，答えが全員一致したデータをシステムに解析させたところ,表 5 の結果が得られた．表 4 の結果と合わせて見ると，“話し手”と“第三者”の解析は成功しているが，“聞き手”と“人一般” の解析は失敗している．

“人一般”の精度が低かったのは，そもそも特定の人物を指さないという条件を判断するのに，論理的な思考を要するからと考えられる．一般論を述べているのか，特定の状況を説明しているのかを人間は自分の知識と照らし合わせて論理的に解答している．システムに人間の背景知識を全て実装するのは困難であるので，まずは，表層的な情報から一般論か否かを判断する規則を見つけるのが今後の課題である．

表 5　基本的な問題の解析結果

	正答数	総数	正解率
人一般	1	3	33 %
話し手	12	14	86%
聞き手	1	2	50%
第三者	4	5	80%
不明	0	0	NULL
計	18	24	75%

7　おわりに

代名詞「自分」の人称判別システムを作成し，適合率 56.7%の精度が得られた．この数値は同じ条件下における人間の平均精度には及ばないが，基本的な文における解析はできたと言えよう．

現在までは 1 文内の情報のみ利用してきたが，今後は文脈情報も考慮して解析を行う予定である．これによって，文の主語または主題が省略されている場合にも前後の文から推測できるようにする．

参考文献

[1] 生田奈穂・吉本啓: 日本語の再帰代名詞『自分』の先行詞の決定条件，言語処理学会第 8 回年次大会発表論文集，pp.96-99，2002.

[2]吉見毅彦: 英日機械翻訳における代名詞翻訳の改良，自然言語処理，Vol.8，No.3，pp.87-106，2001.

[3] 松本裕治，他:形態素解析システム『茶筅』version 2.3.3 使用説明書，2003. http://chasen.naist.jp/hiki/ChaSen/

[4] Taku Kudo, Yuji Matsumoto: Fast Methods for Kernel-Based Text Analysis, ACL, 2003.

[5] Kanako Komiya, Yasuhiro Tajima, Nobuo Inui, Yoshiyuki Kotani, Generating a Set of Rules to Determine Honorific Expression Using Decision Tree Learning, Lecture Notes in Computer Science, Volume 3878, pp.315-318, 2006.

[6]J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc, 1993.